Wearable video monitoring of people with age dementia: Video indexing at the service of helthcare

Rémi Mégret, Daniel Szolgay, Jenny Benois-Pineau, Philippe Joly, Julien Pinquier, Jean-François Dartigues, Catherine Helmer

WEARABLE VIDEO MONITORING OF PEOPLE WITH AGE DEMENTIA : VIDEO INDEXING AT THE SERVICE OF HELTHCARE

R. Megret[(1)], D. Szolgay[(2)], J.Benois-Pineau[(2)], Ph. Joly[(3)], J. Pinquier[(3)], J.-F. Dartigues[(4)], C. Helmer[(4)]

(1) IMS, UMR 5218 CNRS, University of Bordeaux
(2) LaBRI, UMR 5800 CNRS, University of Bordeaux
(3) IRIT, UMR 5505 CNRS, University Paul Sabatier
(4) ISPED, INSERM U593, University of Bordeaux

ABSTRACT

Between 100 and 150 words
attester

1. INTRODUCTION

Exploration of video surveillance material for the purposes of healthcare and home assistance to elderly population is now becoming in the focus of attention of multi-disciplinary research, both medical practitioners and computer scientists. With the ageing of population in the world, the care of dementia diseases becomes one of the medical priorities. Impairment in Instrumental Activities of Daily Living (IADL) related to cognitive decline is a major diagnostic criterion for Dementia [DSM87]. However a valid evaluation of this criteria is often difficult to obtain because of deny or anosognosia by the patient, or by his caregiver [Dar05]. Moreover, this impairment is also considered as a major consequence of the disease and is one the main outcome in prognosis studies or controlled clinical trial for Dementia or Alzheimer's Disease [Hel06]. Thus, an objective and operational tool designed to evaluate IADL functions could be of major interest in clinical or epidemiological studies on Dementia.

In this paper we propose a general framework and specific tool of multimedia indexing for monitoring the daily activity of people with dementia using wearable video cameras. The paper is organized as follows. In Section 2 we describe the design of the acquisition set-up and compare it with existing solutions of wearable monitoring. In Section 3 we formulate and design first solutions for the content indexing tasks both on a single media (video) and on mixed media (audio and video). In Section 4 we summarize first achievements and outline perspectives.

2. DESIGN OF ACQUISITION SET-UP

2.1. Needs and constraints

The purpose of the system is to record the activity of the patient during an extended time interval (from 3 hours for the observation of a specific activity during the day, such as the lunch and its preparation, to 16 hours when a full day is captured without any intervention).

The goal of this system is to enable specialists of age dementia to analyze the behavior of the patients at home, during their usual activities. For this reason, the focus is put on capturing audio and video data at standard video rate (25Hz), in order to convey most of the useful information concerning inter-personal interactions (through speech and physical contact), individual activities (manual or contemplative activities), or the reaction of the person to its environment.

The footprint of the device in term of size and weight must be the lowest possible, considering that it is meant to be worn by aged persons. The design has therefore to be tailored for this target, which is rather constraining due to the possible articulation and muscle disabilities such persons may have. One objective is to produce a device that weights globally less than 500g, which most of the weight being hold in an ergonomic manner, so that it does not induce fatigue after several hours.

2.2. Review of existing similar devices

The MIThril wearable platform [Dev03] is a modular system based on a PDA computer, which allows both the online capture and processing of various signals, such as ECG, blood pressure or inertial motion, which are needed in different medical contexts. The autonomy is limited both by the consumption of the PDA and the storage capacity.

Our medical video capture problematic is close to the setup proposed by the Sensecam project [Hod06], [O'Con07] which also considers a wearable device with a camera and microphone. It was used in a medical context as a retrospective memory aid for people with memory impairment [Ber07]. The Sensecam device captures images at a low frame-rate (of the order of one image every 30 seconds), which is enough to produce a video log of the day, but not enough to be able to properly analyze details on the behavior of the person, which is of concern in our case.

The previous art most relevant to our context has been developed for the WearCam project [Nor07], which uses a camera strapped on the head on young children. This setup allows capturing the field of view of the child together with their gaze with autonomy of one hour. We discuss this choice in the next section.
Another embedded camera setup was proposed the StartleCam project [Hea98], which used a wearable computer as an acquisition device. The weight being an important point in our context, we prefered the WearCam approach of wireless video transmission.

2.3. Camera field of view

The camera position is closely related to which content must be captured. We can identify two main contents of interest for the analysis of the behavior (see fig 1): close field (physical interaction with the environment…), interpersonal environment (where other persons stay when dialoging), and far environment (which contains events that may come up unexpectedly, such as a phone ringing…).
One of the main advantages of using a wearable camera is the ability to capture hands close to the body, such as when manipulating a tool or holding objects, which can bring information on the ability of the person with respect to manipulation. To obtain such observations with fixed cameras, several of them would need to be installed because of the occultation by the body.

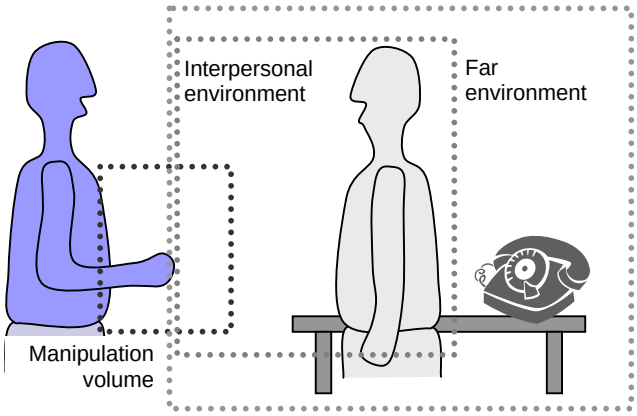

Fig 1. Definition of close, medium and far environment.

In the Sensecam project the camera is placed on the chest, which allows to capture the environment facing the person, at the expense of observing the person's hands (Fig. 2-a).
The camera worn on the head [Nor07] is very good for capturing the intentions and focuses of the person (Fig 2-d), and would certainly be of much interest in our context. However, such an approach was not retained for our first prototype for several reasons. First, the ergotherapists we consulted consider any weight on the head may be a problem for aged persons, because of potential disabilities at the cervical level. Second, the device is intended to be worn in an autonomous setup. The battery that is needed to power the camera during several hours is heavy enough to require it to be worn at some practical location such as the hip. The wires running up to the head would make the device difficult to put and to remove when the person feels the need to.
Two setups have been tested in our case: a first setup is located close to the manipulation area (Fig 2-b), and second one on the shoulder (Fig 2-c), similar to the one that was shown in [May02] to offer a good compromise considering several criteria.
The system being designed to be removable whenever the person feels the need to, the camera position can be displaced by the person themselves, resulting in small angle variations. For that reason, a too strict set-up of the position of the camera with respect to the person can not be kept over a long period, which makes a large field of view camera necessary to be able to cope with such variations.

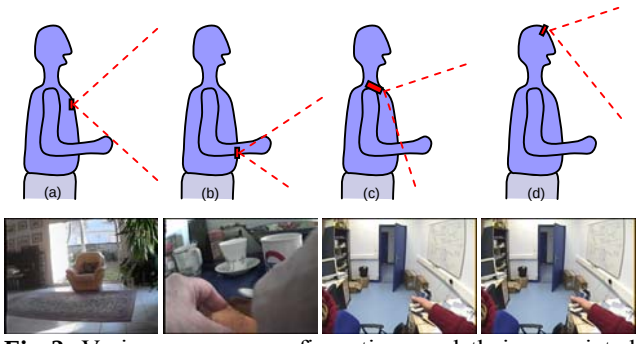

Fig 2. Various camera configurations and their associated field of view.

2.4. Device design and characteristics

Fig 3. Illustration of the current prototype, with considered camera position (a,b,c). The system is composed of a strap bag which holds a battery, a camera, a microphone, and a wireless transmitter.

The monitored person is equipped with an onboard camera and microphone which are integrated into a bag that is attached to the shoulder and the hip. The video and audio signals are transmitted wirelessly to a base recording station via an analog 2.4GHz transmitter within a 100m range, which is enough for capturing the actions inside a house. The camera and radio transmitter weight around 80g, and the 6800mAh battery weights 300g. The battery weight can be reduced when only short observations are needed.

The recording station receives the analog signal, digitizes and compresses it through an acquisition card (USB GrabeeX deluxe), and stores the compressed video on a USB external hard drive. This removable storage with a capacity of 500GB can hold several days of acquisition. In case the person goes away from home, the recording station can be replaced by a portable media recorder extended with a wireless receiver and extra batteries, carried in a bag.

At the end of the observation period, the removable hard drive is easily recovered by a member of the project, and its contents transferred to a centralized and secure archive server. The members of the project can then have access to the videos through a secure network connection, in order for the doctors to make clinical observations of the behavior of the person. The video are also shared to test new algorithms and assess the feasibility of automatic processing that can then be integrated into the video consultation system.

3. CONTENT INDEXING TASKS

To efficiently and automatically index the content, thus proposing browsing facilities and (semi-)automatic interpretation of patient behavior to the doctors, several indexing problems have to be solved. The first straightforward tasks are summarizing video content in order t give a scene overview, identifying specific actions (cooking, watching TV, ...), and identifying interpersonal interaction. These tasks can be solved both in a mono-media framework (video) and by using audio and visual cues togerther. In this paper we address such tasks as scene overview for fast browsing, important inter-personal issue such as privacy check and person detection, and finally the identification of the audio environment.

3.1. Scene Overview

The registered video sequences will be used by medical practiciens, who are interested in particular events (cooking, washing, reading) in particular places (kitchen, garden, living room) in the household and also in outdoor environment. To make their work easier the sequence parts have to be indexed and summarized. To achieve this goal we intend to use mosaic images to sum up the scenes. The mosaic images proved to be an efficient tool for video summarizing and recognition of the environment [Ira98]. In this work we resort to the mosaicing method we developed for the MPEG2 compressed general purpose video[Kra06, Kra07].

The mosaic image of a video segment is a blend of all frames of the segment aligned in the same coordinate system of a reference frame as

$$M = \beta_K \sum_{k=1}^{K} G(k).I(k), \qquad (1)$$

Where $I(k)$ is the kth video frame, $G(k)$ is the geometric transformation from the image $I(k)$ to the mosaic M, K is the number of video frames $I(k)$ in the sequence, and $\beta_K(n)=1/|n|$ with $|n|$ as the number of available pixels at position n.

Thus to compute the mosaic M we need to compute the geometrical transformation $G(k)$ of each video frame $I(k)$ into the reference frame $I(0)$. In [Kra07] this geometrical transformation has been defined as a composition of affine 1st order models of camera motion between successive frames.

$$\vec{d}(c_x, c_y) = \begin{pmatrix} a_1 \\ a_4 \end{pmatrix} + \begin{pmatrix} a_2 & a_3 \\ a_5 & a_6 \end{pmatrix} \begin{pmatrix} c_x \\ c_y \end{pmatrix}, \qquad (2)$$

Here $d(c_x,c_y)$ is a displacement vector of a center of a block of pixels. In [Kra07] we used MPEG2 macro block motion vectors in order to derive the camera model.

Despite the videos acquired with the wearable camera are available in MPEG2 compressed form, the macro-block motion vectors cannot be directly used as it was the case in [Kra07]. This is due to the fact that in the most cases the motion between successive frames has a very strong magnitude. Indeed, the elderly persons often have problems with balance control during walking. Thus the displacement becomes very strong, we observed magnitudes of around 30 pixels in average (see Fig. 1).

Fig. 4.: Three consecutive frames from an outdoor video. The pictures are blurred because of the strong motion.

Thus to correctly align video frames we developed an hierarchical block-matching algorithm and applied it to the decoded frames. Here, for the pair of consecutive frames *I(k-1), I(k)* Gaussian multi-resolution pyramids are built *P(k-1,l),P(k, l) , l=0,..,2* with *l* denoting the level in the pyramid.
The block-matching is realized at the top level of the pyramid by a full-search estimator with a pixel accuracy.
Then the motion vectors of blocks at the nearest lower level of the pyramid $d^{l-1}\left(c_x^{l-1}, c_y^{l-1}\right)$ are predicted as

$$\tilde{d}^{\,l-1}\left(c_x^{l-1}, c_y^{l-1}\right) = s \cdot d^{l}\left(c_x^{l}, c_y^{l}\right) \qquad (3)$$

where *s* is the sub sampling factor: in this work we used a conventional value *s=2*.
Then the motion vectors are refined by a full search with a half-pixel accuracy. This process is repeated until the final motion vectors at the level *l=0* are obtained. This is a classical hierarchical block-matching algorithm used in video compression. Here the window size *W* for the displacement search was chosen sufficiently large at the top-level in order to re-cover strong motion magnitudes (*W=8*) and was significantly reduced at the intermediate levels and the basis of the pyramid (*W=3*).
The resulting motion vectors were used as initial measures for a robust motion estimator [Dur01] allowing for rejection of outliers – and obtaining the global camera model (2). The geometrical transformation *G(k)* in (2) has then been computed as a composition *G(k)=G(k-1)°G(k-2)°...G(1)*.
The example of the mosaic obtained is shown in Fig. 5. Here we can state that the final mosaic is not very much focused. The main reason of this is the low quality of original frames acquired with a wearable camera and exhibiting a strong motion blur. This aspect can be improved in the future by our super-resolution method [Kra07]. Still, this mosaic gives a good overview and shortens a lot the browsing of the content by doctors.

Fig.5: First results of the mosaicing algorithm with 10 pictures.

3.2 Privacy Check

Giving the records to the doctors rise a new problem: there might be recognizable persons on the pictures and this violates their right to privacy. To solve this problem we want to hide the faces of the people who appear on the recorded sequences. This requires finding humans in the video.
Assuming that humans are moving, we search for camera independent motions. Let *I(k-1), I(k)* be two consecutive frames. With (2) we transform *I(k-1)* according to the camera motion between the two frames *I(k-1), I(k).* We use this transformed image, $\tilde{I}$ *(k-1)* to calculate an error image *E(k)*, which shows only those objects that are moving independently from the camera:

$$E(k) = \left|\tilde{I}(k-1) - I(k)\right| \qquad (4)$$

We consider these objects as foreground objects, and the rest of the frame as the background.
Although we estimate the camera motion, the error image always contains pixels that are belonging to background objects. One of the most important thing to do in the future is to make the camera motion estimation as accurate as possible.

Fig. 6: Top Left: Previous Frame(I(k-1)), Top Right: Current Frame (I(k)), Bottom Left: Compensated Frame ($\tilde{I}$ (k-1)), Bottom Right: Error Image (E(k)).

If the compensation gives us a reliable result, the further search will have to be made only on the foreground. We are planning to use a whole body detector on the foreground objects. The problem with these detectors is that they are sensitive to the camera orientation. Usually it is not a serious problem because the camera is positioned uprightly, but in case of wearable cameras like [Nor07] and like ours, the camera can tilt. Usually this results approximately ±45° deflection from the vertical/horizontal direction in the worst case.

To determine the orientation of the camera a simple idea was used. We calculate the edge orientation histogram *H(k)* of each frame *I(k)*. In this work we used 32 bin histograms for the angle range 0-180°. Generally a picture showing an ordinary living environment (house, garden) contains mainly horizontal and vertical edges. This means that we will have one or two peaks in the histogram, and these peaks will determine the horizontal or vertical direction. An exemple is given in Figure 7.

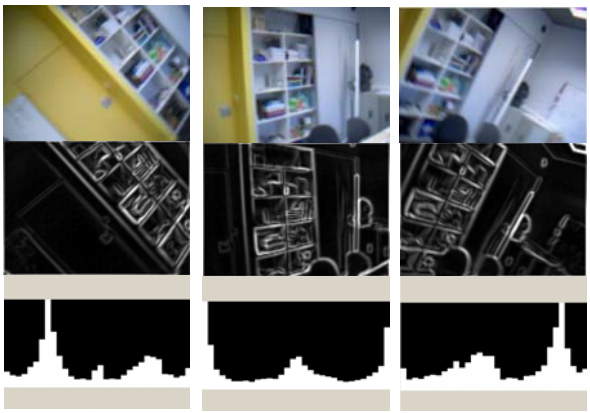

Fig. 7: Changes in the edge orientation histogram changes along with the camera orientation.

What we cannot know from this information is that a peak corresponds to the vertical or the horizontal direction. But in our case it is not a real problem because we can assume that the real vertical/horizontal direction in the ground thruth scene is closer to the vertical/horizontal direction of the scene registered by the camera. With this assumption the result of the algorithm is unambiguous, however it can only handle angular offsets (α) that are smaller then 45°.

Later we would like to improve this method by using the camera motion information to make the results more reliable.
Knowing the angular offset α, the image can be rotated to upright position with a geometrical rotation around the center of the image. The matrix of the rotation:

$$R(\alpha)=\begin{pmatrix}\cos\alpha & -\sin\alpha\\ \sin\alpha & \cos\alpha\end{pmatrix}$$

where α is the angular offset of the camera. To test the effectiveness of our algorithm we used Haar-like feature based face detector ([Vio01] ,[Lien02]). An exemple of geometrical correction proposed and face detection is shown in figure 8.

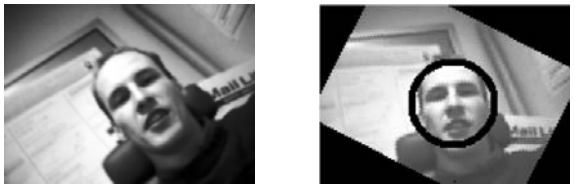

Fig. 8: Face detection after geometrical correction. -On original frame (left) the face detector fails to detect the face, while on the rotated image (right) it detects the face.

In the future we would like to use a whole body detector instead of face detector. This is necessary because in some cases the face detection is not possible (partly visible or too small faces), but we need to be aware of the presence of other persons.

3.3 Audiovisual indexing tools

3.3.1 Tools presentation

In regards with the identified events which may be of interest for the medical domain, some tools developed for cultural audiovisual content indexing were considered to extract potential key features for further works.

We first evaluated the interest of applying a segmentation tool on the audio track aiming to localize speech, music noise and silent segments. Speech segment can be useful to index conversational situations whereas music may help to identify activities such as watching TV or listening to the radio. This tool is divided into two classifications (speech/non-speech and music/non-music) [Pin02]. For speech detection, we use an original feature (entropy modulation) and we merge it with the classical 4 Hertz modulation energy (corresponds to the syllabic rate). For music detection, we extract other original parameters: number of segments and segment duration. There result from the FBD algorithm [And88]. The decision is made regarding to the maximum likelihood criterion (scores). Finally, we have two classifications for each second of input signal: the speech/non-speech one and the music/non-music one. Then, we can merge them to have a speech/music classification system composed of four different segments: Speech, Music, Speech-Music and none (silence or noise). System provides more than 90% of accuracy for speech detection and music detection on audiovisual documents (like TV and Radio broadcasts).

The second tool applied on the first recordings made in the framework of the project consists in detecting faces in the video, and labeling persons on the base of their clothes appearance. The idea would be here to localize phases of interaction with the relatives of the patient or with unknown

people. This tool detects faces using the face detector available in OpenCV [Vio03, Op]. Once a face has been detected at the same place on a sufficient number of frames, color and texture features are extracted from the zone located below, which is supposed to correspond to the person's clothes. These descriptors are then compared with those which may already have been extracted on previous images. If it appears that these values are met for the first time, the instance of a new person is created in the database. In the other case, a new occurrence of an already known person is inferred [Jaf05]. As far as this tool was developed for television applications, some priors have been used to improve results. The first one takes into account the fact that a same person is more likely to appear in the same set. This means that some color features related with the background have been added in the signature describing a person in the database. The second prior leads to consider that a person is always present on the screen for a minimal duration. For example, inferring the detection of a person during a whole shot, even if he or she was actually detected on a few frames, leads to significant improvements of the method on television programs. Unfortunately, this second prior can not be applied in the paradigm of the proposed experiments. We have so decided that a person will be detected only if he or she appears on seven successive frames (this threshold has been determined as an optimal one in the general case). The output of that tool is a list of labels attached to the detected persons appearing in the video and, for each of them, a list of temporal segments corresponding to their appearances.

3.3.2. Application

In order to identify the possibility to exploit these tools and their limitations in this context, they were applied on 6 minutes extracted from two different recordings. The first one is an indoor recording showing during the first minutes three different persons. It has been shot with a static camera in the kitchen environment The audio track is very noisy while the video is quite clear. Speech on the soundtrack can hardly be understood. The second recording has been taken outside, in a garden. The patient is doing gardening activities. A relative can be seen on the first minute of video. Here, the audio track is much clearer than in the first case, but in the same time, the video is much more difficult to process: there are fast motions with a high magnitude corresponding to objects (hands, tools) close to the camera, and camera motions are producing really instable images. There is a lot of noise due to the poor quality of the wireless communication between the remote camera and computer with is located in the house. At last, of course, framing is anything but on purpose and people appearances are most of the time truncated (the head is most of the time out of the field of the camera).

3.3.3. Analysis of results

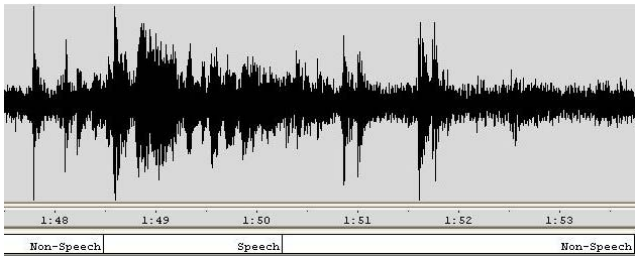

Figure 9: Example of manual speech detection during a few seconds.

The performance of the speech/music detection algorithm is about 75%: this is very low compared to a TV broadcast corpus. If we only focus on the person wearing the microphone, the score increased by more than 10%. In the figure 9, we present an example of manual speech detection to highlight difficulties inherent to the corpus. The signal is very noisy, even in areas of non-speech (which are supposed to be silent segments!). In this example, our system has detected no speech during these few seconds.

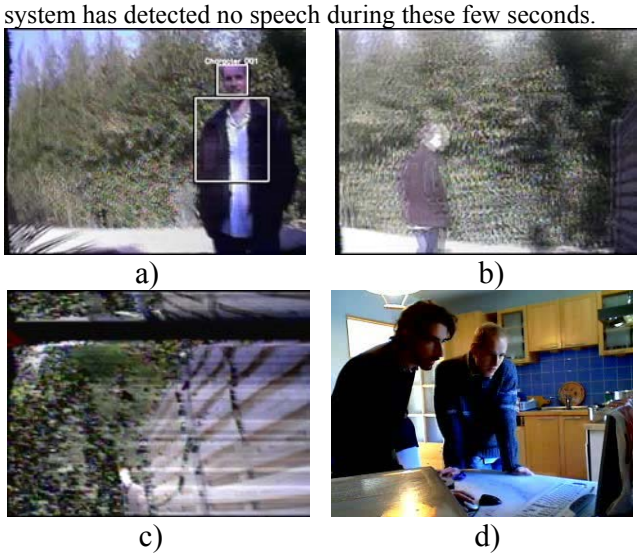

Figure 10: a) example of a person detected in the "outdoor" excerpt; b) and c) examples of noisy frames surrounding this detection; d) persons not detected in the "inside" excerpt.

Performances of the person detector and labeling are very low compared with results which can be obtained on classical TV programs (where this tool detects persons with

recall and precision rates both of about 97%). In the case of those excerpts, the tool does not produce any false positive, but the recall rate is about 5%...
Surprisingly, the best results were obtained on the noisiest excerpt (the one corresponding to the outside shot). The lighting conditions for the "inside excerpt" are actually too low for the face detector.

Further works will have to overcome the observed limitations of these tools. For the soundtrack segmentation and classification tool, we have to solve different problems: reducing the noisiness of audio environment and detecting activity areas.
For the first one, indeed noise is highly variable: it can be transmission breaks of information, but also to strong variations of the environment that influence audio recordings. Noise is not constant but depends on the person activity and the place in which it is located. This step is a necessary pre-processing for audio detection.
For the second one, it could be interesting to identify recurring speakers (such as family members), specific locations (kitchen) or specific items (phone, TV). It will be necessary to increase our "audio classes" (speech and music) with other key sounds (phone ringing, kitchen tools, speakers, etc.) and to detect the sound environment.

For the visual detection and identification of person, the first idea to be explored will be the usage of new priors. The first one will be to consider that the number of persons which may appear is strongly limited. As far as this system is supposed to be used only at home or in the close proximity of the patient's house, only the family members and a few friends are expected to appear on the recordings. Furthermore, on the same recording, we can assume that a same person is wearing the same clothes. These priors can be used to improve both the detection and the identification process because they allow to maximize the number of classes. So, clothes descriptors associated to reliable detections of faces could then be used to directly detect clothes themselves. In this case, face detection will not necessary be a trigger for this tool and people could then be detected even if they not fully appear on the screen. This may solve the framing problem which has been identified in these preliminary works. The "outside excerpt" shows that the persons can be seen generally during very short segments before the camera moves to another point of view. So, the second point will be to modify the prior about the minimal duration of a face appearance on the screen used to confirm the person detection.

4. FIRST RESULTS AND PERSPECTIVES

The project of wearable video monitoring has fro people with dementia is the first application of video indexing methods in the studies of this disease. This is why the closest collaboration between the information technology team and medical research team was a must. Due to the specification of the content of interest and the specific ergonomics for aged patients, the experimental acquisition set-up has been designed en the first corpus registered. At the present stat it contains 7 hours 29 minutes video from a wearable camera. These records were taken with a healthy, elderly person, whose actions were not restricted in any way during the recordings. She was doing her everyday routine in the house and in the garden. Examples of key –frames of a daily activity are shown in fig. 11

We have also 2.5 hours minutes video from a static camera, that was placed in a strategic point (kitchen) in the house. In the future we intend to use the static camera along with the wearable camera in full time.

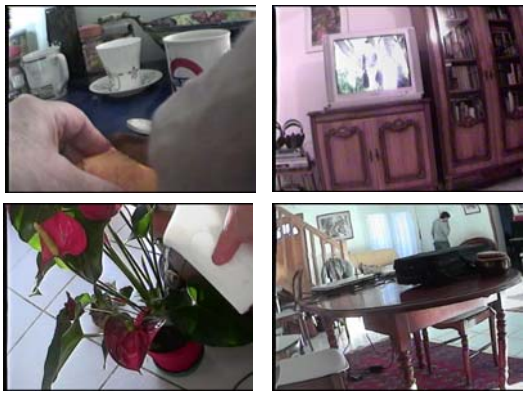

Figure 11. Key-frames of a daily activity

In order to get the continuous feed-back form patricians, an video browsing interface has been developed allowing for navigation according to the time –line and to the activities.
It is presented in Fig. 12. In the future it will serve as a front-end for the automatic analyses tools we proposed and reported in Section 3 and further methods.

5. CONCLUSION

Hence in this paper we proposed a general framework, designed the acquisitions set-up and developed first methods for wearable monitoring of patients with dementia.

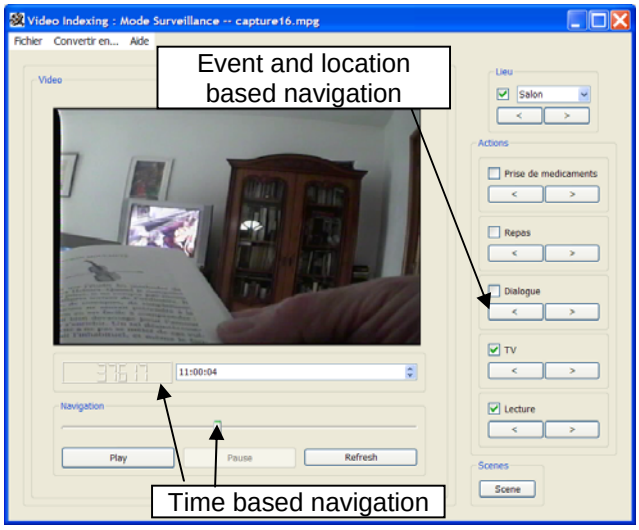

Fig. 12. Screenshot of the video indexing and navigation interface.

6. REFERENCES

[DSM87]American Psychiatric Association. Diagnostic and statistical manual of Mental Disorders. DSM III-R. Washington DC: Amer Psychiatr Ass; 1987

[Dar05]Dartigues JF. [Methodological problems in clinical and epidemiological research on ageing]. Rev Epidemiol Sante Publique 2005;53(3):243-9.

[Hel06]Helmer C, Peres K, Letenneur L, Guttierez-Robledo LM, Ramaroson H, Barberger-Gateau P, Fabrigoule C, Orgogozo JM, Dartigues JF. Dementia in subjects aged 75 years or over within the PAQUID cohort: prevalence and burden by severity. Dement Geriatr Cogn Disord 2006;22(1):87-94.

[Dev03] R. DeVaul, M. Sung, J. Gips, and A. Pentland, "MIThril 2003: Applications and Architecture," in *International Symposium on Wearable Computers (ISWC)*. IEEE, 2003, pp. 4-11.

[Ber07] E. Berry, N. Kapur, L. Williams, S. Hodges, P. Watson, G. Smyth, J. Srinivasan, R. Smith, B. Wilson, and K. Wood, "The use of a wearable camera, SenseCam, as a pictorial diary to improve autobiographical memory in a patient with limbic encephalitis," *Neuropsychological Rehabilitation*, vol. 17, numbers 4/5, pp. 582-681, August 2007.

[Hod06] S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur and K. Wood, "SenseCam: a Retrospective Memory Aid," in *International Conference on Ubiquitous Computing (UBICOMP)*. 2006, LNCS 4206, pp. 177-193.

[O'Con07]O'Conaire C, O'Connor N, Smeaton A.F. and Jones G. Organising a daily Visual Diary Using Multi-Feature Clustering, SPIE Electronic Imaging - Multimedia Content Access: Algorithms and Systems (EI121)/, San Jose, CA, 28 January - 1 February 2007.

[Hea98] Jennifer Healey, Rosalind W. Picard, "StartleCam: A Cybernetic Wearable Camera," in *International Symposium on Wearable Computers (ISWC)*. IEEE, 1998, pp. 42-49.

[Pic07] L. Piccardi, B. Noris, O. Barbey, A. Billard, G. Schiavone, F. Keller and C. von Hofsten, "WearCam: A head mounted wireless camera for monitoring gaze attention and for the diagnosis of developmental disorders in young children," in *International Symposium on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2007, pp. 594-598.

[May02] W.W. Mayol, B. Tordoff and D.W. Murray, "Designing a Miniature Wearable Visual Robot," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2002, vol. 4, pp. 3725-3730.

[Ira98] M. Irani and P. Anandan "*Video indexing based on mosaic representation*". In Proc of the IEEE, vol.86,1998

[Kra06] P. Kraemer, O. Hadar, J. Benois-Pineau, J.-P. Domenger "Use of Motion Information in *Super-Resolution Mosaicing,*" Proc. IEEE International Conference on Image Processing (ICIP), pp 357-360, 2006.

[Kra07] P. Kraemer, O. Hadar, J. Benois-Pineau, J.-P. Domenger *"Super-Resolution Mosaicing from MPEG Compressed Video,"* Signal Processing: Image Communication, Volume 22 ,Issue 10, pp 845-865, 2007.

[Dur01] M. Durik, J. Benois-Pineau, *"Robust motion characte-risation for video indexing based on MPEG2 opticalflow,"* Proceedings of the International Workshop on Content-Based Multimedia Indexing, CBMI'01, pp. 57–64, 2001.

[Lie02] Rainer Lienhart and Jochen Maydt, *"An Extended Set of Haar-like Features for Rapid Object Detection,"* IEEE ICIP 2002, Vol. 1, pp. 900-903, Sep. 2002.

[Vio01] Paul Viola and Michael J. Jones. *"Rapid Object Detection using a Boosted Cascade of Simple Features,"* IEEE CVPR, 2001.

[Pin02] Pinquier, J., J.-L. Rouas, and André-Obrecht, R., "Robust speech / music classification in audio documents". In: *International Conference on Spoken Language Processing*, Vol. 3. Denver, USA, pp. 2005–2008, May 2002.

[And88] André-Obrecht, R., "A New Statistical Approach for Automatic Speech Segmentation". *IEEE Transactions on Audio, Speech, and Signal Processing*, 36(1), 1988.

[Op] OpenCV: http://www.intel.com/research/mrl/research/opencv/

[Jaf05] Jaffré, G. and Joly, P., "Improvement of a Temporal Video Index Produced by an Object Detector". In: *Proceedings of the 11th International Conference on Computer Analysis of Images and Patterns*, Rocquencourt, France, September 2005.